massachusetts institute of technology — artificial intelligence laboratory

On the Dirichlet Prior and Bayesian Regularization

Harald Steck and Tommi S. Jaakkola

AI Memo 2002-014 September 2002

Abstract

A common objective in learning a model from data is to recover its network *structure*, while the model parameters are of minor interest. For example, we may wish to recover regulatory networks from high-throughput data sources. In this paper we examine how Bayesian regularization using a Dirichlet prior over the model parameters affects the learned model *structure* in a domain with discrete variables. Surprisingly, a weak prior in the sense of smaller equivalent sample size leads to a strong regularization of the model structure (sparse graph) given a sufficiently large data set. In particular, the *empty* graph is obtained in the limit of a vanishing strength of prior belief. This is diametrically opposite to what one may expect in this limit, namely the complete graph from an (unregularized) maximum likelihood estimate. Since the prior affects the parameters as expected, the prior strength balances a "trade-off" between regularizing the parameters or the structure of the model. We demonstrate the benefits of optimizing this trade-off in the sense of predictive accuracy.

This work was supported by the German Research Foundation (DFG) under grant STE 1045/1-1, Nippon Telegraph and Telephone Corporation, and NSF ITR grant IIS-0085836.

1 Introduction

Regularization is essential when learning from finite data sets. In the Bayesian approach, regularization is achieved by specifying a prior distribution over the parameters and subsequently *averaging* over the posterior distribution. This regularization provides not only smoother estimates of the parameters compared to maximum likelihood but also guides the selection of model structures.

It was pointed out in [9] that a very strong prior belief can degrade predictive accuracy due to severe regularization of the *parameter* estimates. We complement this discussion here and show that a very *weak* prior belief can lead to poor over-regularized *structures* when the (conjugate) *Dirichlet prior* is used over multinomial conditional distributions (Section 3). Section 4 demonstrates the effect of the strength of prior belief and how it can be calibrated. We focus on the class of Bayesian network models throughout the paper.

2 Regularization of Parameters

We briefly review Bayesian regularization of *parameters*. We follow the assumptions outlined in [9]: multinomial sample, complete data, parameter modularity, parameter independence, and Dirichlet prior. Note that the Dirichlet prior over the parameters is often used for two reasons: (1) the conjugate prior permits analytical calculations, and (2) the Dirichlet prior is intimately tied to the desirable likelihood-equivalence property of network structures [9]. The Dirichlet prior over the parameters $\theta_{\cdot|\pi_i}$ is given by

$$p(\theta_{x_i|\pi_i}) = \frac{\Gamma(\sum_{x_i} \alpha_{x_i,\pi_i})}{\prod_{x_i} \Gamma(\alpha_{x_i,\pi_i})} \; \theta_{x_i|\pi_i}^{\alpha_{x_i,\pi_i}-1}, \tag{1}$$

where $\theta_{x_i|\pi_i}$ pertains to variable X_i in state x_i given that its parents Π_i are in joint state π_i ($i = 1, ..., n$; n is the number of variables in the domain). The normalization terms in Eq. 1 involve the Gamma function $\Gamma(\cdot)$. The positive hyper-parameters α_{x_i,π_i} of the Dirichlet distribution can be interpreted as *pseudo-counts* implied by the prior belief. This is apparent when calculating the *average* parameter value, which typically serves as the regularized parameter estimate given a network structure m,

$$\bar{\theta}_{x_i|\pi_i} \equiv E_{p(\theta_{x_i|\pi_i}|D,m)}[\theta_{x_i|\pi_i}] = \frac{N_{x_i,\pi_i} + \alpha_{x_i,\pi_i}}{N_{\pi_i} + \alpha_{\pi_i}}, \tag{2}$$

where N_{x_i,π_i} are the cell-counts from data D; $E[\cdot]$ is the expectation.

There are a number of approaches to specifying the hyper-parameters α_{x_i,π_i} [4, 3, 9]; the common choice,

$$\alpha_{x_i,\pi_i} = \alpha \cdot p(x_i, \pi_i), \tag{3}$$

where p is a (marginal) prior distribution of pseudo-counts, ensures likelihood equivalence of the network structures [9]. If p is chosen to be uniform, $p(x_i, \pi_i) =$

$1/|X_i|/|\Pi_i|$, where $|\cdot|$ denotes the number of (joint) states, an uninformative prior is obtained [3]. The hyper-parameter α is independent of X_i, and it is typically called the *equivalent sample size*. Since the equivalent sample size represents the overall number of pseudo-counts, $\alpha = \sum_{x_i,\pi_i} \alpha_{x_i,\pi_i}$, it may also be thought of as the *strength* of prior belief. The assignment according to Eq. 3 is assumed throughout this paper.

The pseudo-counts lead to regularized parameter estimates: the estimated parameters become "smoother" or "less extreme", because the assumed prior distribution p typically is rather smooth, i.e., close to uniform. An increasing strength of prior belief, α, leads to a stronger regularization, i.e., to "smoother" parameter estimates (cf. Eq. 2); in the limit of a vanishing prior strength, $\alpha \to 0$, the (unregularized) maximum likelihood estimate is obtained, as expected.

3 Regularization of Structure

In the remainder of this paper, we outline surprising effects due to Bayesian regularization of the Bayesian network structure when using the Dirichlet prior. Let us briefly introduce relevant notation.

In the Bayesian approach to structural learning, the posterior probability of the network structure m is given by $p(m|D) = p(D|m)p(m)/p(D)$, where $p(m)$ denotes the prior distribution over the network structures; $P(D)$ is the (unknown) probability of given data D, and $p(D|m)$ is the marginal likelihood of m. A natural approach to comparing different model structures is to devise so-called *relative scores*, which be defined as the difference in the (log) scores of two structures m^+ and m^-:

$$\log \frac{p(m^+|D)}{p(m^-|D)} \quad = \quad \log \frac{p(m^+)}{p(m^-)} + \log \frac{p(D|m^+)}{p(D|m^-)}, \tag{4}$$

where the posterior ratio decomposes into the prior ratio and the Bayes factor. In particular, let us consider two graphs that are identical except for a single edge, say $A \leftarrow B$ between the variables A and B. Let the edge be present in graph m^+ and absent in m^-. Adopting the assumptions outlined in [9], including the Dirichlet prior over the parameters θ, the marginal likelihood $p(D|m) = E_{p(\theta|m)}[p(D|m,\theta)]$ can be calculated analytically, and one arrives at the log Bayes factor (cf. also [4]):

$$\begin{aligned} \log \frac{p(D|m^+)}{p(D|m^-)} \quad = \quad & \sum_{a,b,\pi} \log \frac{\Gamma(N_{a,b,\pi} + \alpha_{a,b,\pi})}{\Gamma(\alpha_{a,b,\pi})} - \sum_{a,\pi} \log \frac{\Gamma(N_{a,\pi} + \alpha_{a,\pi})}{\Gamma(\alpha_{a,\pi})} \\ & - \sum_{b,\pi} \log \frac{\Gamma(N_{b,\pi} + \alpha_{b,\pi})}{\Gamma(\alpha_{b,\pi})} + \sum_{\pi} \log \frac{\Gamma(N_{\pi} + \alpha_{\pi})}{\Gamma(\alpha_{\pi})} \,. \end{aligned} \tag{5}$$

The notation $\pi = \pi_A$ is used for brevity. Note that the summation involves only the (joint) states a, b and π of the variables A, B and the parents Π_A. The terms pertaining to the remaining variables in the network structures m^+ and

m^- cancel out, as the marginal likelihood decomposes under the assumptions outlined in [9].

A positive value of the relative score indicates that the presence of the edge $A \leftarrow B$ is favored, given the parents Π_A; conversely, a negative relative score suggests that the absence of this edge is preferred.

3.1 Limit of Vanishing Equivalent Sample Size

This section is concerned with the limit of a vanishing strength of prior belief, $\alpha \to 0$. As it turns out, in this limit Bayesian regularization depends crucially on the number of zero-cell-counts in the contingency table implied by the data.

Definition 1: *Let the* Effective Degrees of Freedom (EDF) *be given by*

$$d_{\mathrm{EDF}} \equiv \sum_{a,b,\pi} I(N_{a,b,\pi}) - \sum_{a,\pi} I(N_{a,\pi}) - \sum_{b,\pi} I(N_{b,\pi}) + \sum_{\pi} I(N_{\pi}), \qquad (6)$$

where $N_{a,b,\pi}$, $N_{a,\pi}$, $N_{b,\pi}$, N_{π} are the (marginal) cell counts in the contingency table implied by the data; $I(\cdot)$ is an indicator function such that $I(x) = 0$ if $x = 0$ and $I(x) = 1$ otherwise.

In case of all cell counts being positive, the EDF are identical to the well-known *degrees of freedom* (DF), $d_{\mathrm{EDF}} = d_{\mathrm{DF}} = (|A| - 1)(|B| - 1)|\Pi|$. The EDF / DF play an important role in regularizing the network structure learned from data, see also Appendix A.1. The key difference is that EDF accounts for zero-cell-counts implied by the data.

Let us now consider the behavior of the Bayes factor (cf. Eq. 5) in the limit of small equivalent sample size. We find

Proposition 1: *Let m^+ and m^- be the two network structures as defined in Section 3; let the prior belief be given according to Eq. 3. Then in the limit $\alpha \to 0$:*

$$\log \frac{p(D|m^+)}{p(D|m^-)} \to \begin{cases} -\infty & \text{if } d_{\mathrm{EDF}} > 0, \\ +\infty & \text{if } d_{\mathrm{EDF}} < 0. \end{cases} \qquad (7)$$

The result holds independently of a particular choice of the prior distribution in the pseudo-counts, $p(x_i, \pi_i) > 0$. If the prior over the network structures is strictly positive, this limiting behavior also holds for the posterior ratio (cf. Section 3).

The proof is given in Appendix A.2, together with a brief discussion of the case $d_{\mathrm{EDF}} = 0$. A few comments on Proposition 1 are in order: in the limit of a vanishing strength of the prior, the absence (presence) of the edge $A \leftarrow B$ is favored by the Bayes factor, given positive (negative) EDF. Since this behavior applies to every edge in the network, it follows immediately that the *empty* (*complete*) graph is assigned the highest Bayesian score when EDF are positive (negative). The regularization in the case of positive EDFs is therefore extreme, permitting only the empty graph. This is precisely the opposite of what one may have expected in this limit, namely the complete graph corresponding to

the *unregularized* maximum likelihood estimate (MLE). In contrast, when EDF are negative, the complete graph is favored. This agrees with MLE.

Roughly speaking, positive (negative) EDF correspond to large (small) data sets. It is thus surprising that a small data set, where one might expect an increased restriction on model complexity, actually gives rise to the complete graph, while a large data set yields the most regularized (empty) graph in the limit $\alpha \to 0$. Moreover, it is conceivable that a "medium" sized data set may give rise to *both* positive and negative EDF. This is because the *marginal* contingency tables implied by the data with respect to a sparse (dense) graph may contain a small (large) number of zero-cell-counts. The relative Bayesian score can hence become rather unstable in this case, as completely different graph structures are optimal in the limit $\alpha \to 0$, namely graphs where each variable has either the maximal number of parents or none.

While the *relative* Bayesian score diverges in the limit $\alpha \to 0$, the absolute score, e.g. marginal likelihood, of any graph m vanishes in this limit $\alpha \to 0$. This can be seen in a similar way as in Appendix A.2. The divergence of the relative Bayesian score also implies that there exists a (small) positive threshold value $\alpha_O > 0$ such that, for any $\alpha < \alpha_O$, the same graph(s) are favored as in the limit.

Note that there are two reasons for pseudo-counts α_{x_i,π_i} to take on small values (cf. Eq. 3): (1) a small equivalent sample size α, or (2) a large number of joint states, i.e. $|X_i| \cdot |\Pi| \gg \alpha$, due to a large number of parents (with a large number of states).

The surprising effect noted in Proposition 1 is a consequence of the *Dirichlet* prior distribution. When $\alpha \to 0$, the Dirichlet prior converges to a discrete distribution over the parameter simplex in the sense that the probability mass concentrates on the corners of the simplex containing $\theta_{\cdot|\pi}$. This is due to the vanishing pseudo-counts $\alpha_{x,\pi}$. The counts can also vanish in the limit of a large number of configurations (x, π) even though the equivalent sample size α remains fixed. This is precisely the limit defining Dirichlet processes [6], which, analogously, produce discrete samples. With a finite data set and a large number of joint configurations, only the typical limit in the Proposition is possible. This follows from the fact that a large number of zero-cell-counts forces EDF to be negative. The surprising behavior implied by Proposition 1 therefore does not carry over to Dirichlet processes.

3.2 Large Equivalent Sample Size

In the other limiting case, where $\alpha \to \infty$, the Bayes factor approaches a finite value, which in general depends on the given data and on the prior distribution of pseudo-counts, p. This can be seen easily by applying the Stirling approximation to the Gamma function in the limit $\alpha \to \infty$. When the popular choice of an uninformative prior is used [3], then

$$\log \frac{p(D|m^+)}{p(D|m^-)} \to 0 \quad \text{as} \quad \alpha \to \infty\,, \tag{8}$$

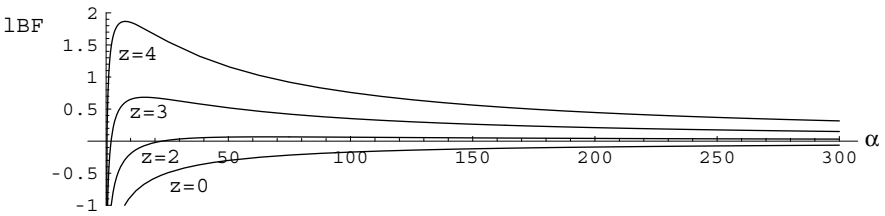

Figure 1: The log Bayes factor (lBF) is depicted as a function of the prior strength α (cf. Eq. 5). It is assumed that the two variables A and B are binary and have no parents; and that the "data" imply the contingency table: $N_{A=0,B=0} = N_{A=1,B=1} = 10 + z$ and $N_{A=1,B=0} = N_{A=0,B=1} = 10 - z$, where z is a free parameter determining the statistical dependence between A and B. An uninformative prior is used.

which is independent of the data. Hence, neither the presence nor the absence of the edge between A and B is favored in this limit.

The behavior of the Bayes factor between the two limits $\alpha \to 0$ and $\alpha \to \infty$ is exemplified for positive EDF in Figure 1: there are two qualitatively different behaviors, depending on the degree of statistical dependence between A and B. A sufficiently weak dependence results in a monotonically increasing Bayes factor which favors the absence of the edge $A \leftarrow B$ at any finite value of α. In contrast, given a sufficiently strong dependence between A and B, the log Bayes factor takes on positive values for all (finite) prior strengths α exceeding a certain value α_+. Moreover, α_+ grows as the statistical dependence between A and B diminishes. This reveals that, in a domain with a range of degrees of statistical dependences, the number of edges in a learned graph can be expected to *increase* with growing prior strength α, as increasingly weaker statistical dependencies between variables are recovered.

This suggests that regularization of network structure *diminishes* with a growing prior strength. Note that this is in the *opposite* direction to the regularization of parameters (cf. Section 2). Hence, the strength α of the prior belief determines the "trade-off" between regularizing the parameters vs. the structure of the Bayesian network model.

If a uniform prior over the network structures is chosen, $p(m)$ =const, the above discussion also holds for the posterior ratio (instead of the Bayes factor). The behavior is more complicated, however, when a non-uniform prior is assumed (cf. Eq. 4). For instance, when a prior is chosen that penalizes the presence of edges, the posterior favors the absence of an edge not only at sufficiently small pseudo-counts, but also at sufficiently large ones. This is apparent from Fig. 1, when the Bayes factor is compared to a *positive* threshold value (instead of zero).

4 Example

This section exemplifies that the *entire* model (parameters *and* structure) has to be considered when learning from data. This is because regularization of model structure diminishes, while regularization of parameters increases with a growing prior strength α, as discussed in the previous sections.

When the entire model is taken into account, one can use a sensitivity analysis in order to determine the dependence of the learned model on the prior strength α, given a prior p defining the pseudo-counts (cf. Eq. 3). The influence of the prior strength α on *predictive* accuracy of the model can be assessed by cross-validation or, in a Bayesian approach, prequential validation [12, 5]. Another possibility is to treat the prior strength α as an additional parameter of the model to be learned from data. Hence, prior belief regarding the parameters θ can then enter only through the (normalized) distribution p over the pseudo-counts. However, note that this is sufficient to determine the (*average*) *prior* parameter estimate $\bar{\theta}$ (cf. Eq. 2), i.e., when $N = 0$. Assuming an (improper) uniform prior distribution over α, its posterior distribution is $p(\alpha|D) \propto p(D|\alpha)$, given data D. Then $\alpha_D = \text{argmax}_\alpha p(D|\alpha) = \sum_m p(D|\alpha, m)\, p(m)$[1] can be calculated exactly if the summation is feasible (like in the example below). Alternatively, assuming that the posterior over α is strongly peaked, the likelihood may also be approximated by summing over the k most likely graphs m only ($k = 1$ in the most extreme case; empirical Bayes). Subsequently, model structure m and parameters $\bar{\theta}$ can be learned with respect to the Bayesian score employing α_D.

In the following, the effect of various prior strengths α is exemplified concerning the data set gathered from Wisconsin high-school students by Sewell and Shah [11]. This domain comprises 5 discrete variables, each with 2 or 4 states; the sample size is 10,318. In this small domain, *exhaustive* search in the space of Bayesian network structures is feasible (29,281 graphs). A uniform prior distribution over the pseudo-counts as well as the graphs is assumed. Figure 2 shows that the number of edges in the graph with the highest posterior probability grows with increasing prior strength, as expected (cf. Section 3). In addition, cross-validation indicates best predictive accuracy of the learned model at $\alpha \approx 100, ..., 300$, while the likelihood $p(D|\alpha)$ takes on its maximum at $\alpha_D \approx 69$. Both approaches agree on the same network structure, which is depicted in Fig. 3. This graph can easily be interpreted in a causal manner, as outlined in [8].[2] We note that this graph was also obtained in [8] due, however, to additional constraints concerning network structure, as a rather small prior strength of $\alpha = 5$ was used. For comparison, Fig. 3 also shows the highest-scoring unconstraint graph due to $\alpha = 5$, which does not permit a causal interpretation, cf. also [8]. This illustrates that the "right" choice of prior strength, when accounting for both model structure and parameters, can have a crucial impact on the learned network *structure* and the resulting insight

[1] We assume that m and α are independent a priori, $p(m|\alpha) = p(m)$.

[2] Since we did not impose any constraints on the network structure, unlike to [8], Markov-equivalence leaves the orientation of the edge between the variables IQ and CP unspecified.

α	a.	XV 5	$\frac{p(D\|\alpha)}{p(D\|\alpha_D)}$
5	6	0.045	10^{-10}
50	7	0.044	0.13
100	7	0.040	0.05
200	7	0.040	10^{-14}
300	7	0.040	10^{-30}
500	7	0.042	10^{-65}
1,000	8	0.047	10^{-151}

Figure 2: As a function of α: number of arcs (a.) in the highest-scoring graph; average KL divergence in 5-fold cross-validation (XV 5), std= 0.006; likelihood of α when treated as an additional model parameter ($\alpha_D = 69$).

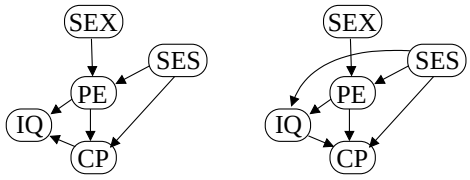

Figure 3: Highest-scoring (unconstraint) graphs when $\alpha = 5$ (left), and when $\alpha = 46, ..., 522$ (right). Note that the latter graph can also be obtained at $\alpha = 5$ when additional constraints are imposed on the structure, cf. [8]. The variables are abbreviated as in [8].

in the ("true") dependencies among the variables in the domain.

Acknowledgments

We would like to thank Chen-Hsiang Yeang for helpful discussions. Harald Steck acknowledges support from the German Research Foundation (DFG) under grant STE 1045/1-1. Tommi Jaakkola acknowledges support from Nippon Telegraph and Telephone Corporation and from NSF ITR grant IIS-0085836.

References

[1] M. Abramowitz and I A. Stegun. *Handbook of Mathematical Functions.* National Bureau of Standards, 1972.

[2] Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis.* MIT Press, 1975.

[3] W. Buntine. Theory refinement on Bayesian networks. *UAI*, pp. 52–60, 1991.

[4] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–47, 1992.

[5] A. P. Dawid. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:277–305, 1984.

[6] T. S. Ferguson. A Bayesian analysis of some nonparamteric problems. *Annals of Statistics*, 1:209–30, 1973.

[7] D. M. A. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–55, 1988.

[8] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pp. 301–54. Kluwer, 1996.

[9] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[10] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–64, 1978.

[11] W. Sewell and V. Shah. Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 73:559–572, 1968.

[12] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–47, 1974.

APPENDIX

A.1 Asymptotic Expansion of the Bayes Factor and EDF

In the following, we give an asymptotic expansion of the Bayes Factor, where we allow for zero-cell-counts N_{x_i,π_i}. In other words, in the limit $N \to \infty$, where N is the sample size, we do not require the asymptotic distribution q ($N_{x_i,\pi_i}/N \to q(x_i,\pi_i)$) to be strictly positive. Applying the Stirling approximation to the Gamma functions in Eq. 5, it follows easily in the limit $N \to \infty$ (for fixed α):

$$\begin{aligned} \log \frac{p(D|m^+)}{p(D|m^-)} &= \sum_{a,b,\pi} N_{a,b,\pi} \log \frac{N_{a,b,\pi} N_\pi}{N_{a,\pi} N_{b,\pi}} - \frac{1}{2} d_{\mathrm{EDF}} \log(N) \\ &\quad - \Big(\sum_{\substack{a,b,\pi: \\ N_{a,b,\pi}=0, \\ N_{a,\pi}, N_{b,\pi}, N_\pi > 0}} - \sum_{\substack{a,b,\pi: \\ N_{a,\pi}=N_{b,\pi}=0, \\ N_\pi > 0}} \Big) \alpha_{a,b,\pi} \log(N) \\ &\quad + \mathcal{O}(1) \end{aligned} \tag{9}$$

This shows explicitly that the (effective) degrees of freedom play a key role in regularizing the learned network structure. As expected, the Bayesian Information Criterion (BIC) [10, 7] is obtained in the absence of zero-cell-counts. However, zero-cell-counts entail a modification of the penalty for model complexity. Eq. 9 shows that, if the data implies zero-cell-counts, the pseudo-counts $\alpha_{a,b,\pi} > 0$ pertaining to the empty cells can have an impact on the Bayes factor even in the asymptotic limit $N \to \infty$. Note the following two particular assignments: (1) if $\alpha_{a,b,\pi} = 1/2$, penalty for model complexity becomes independent of zero-cell-counts, which leads to standard BIC (unless entire rows or columns in the (marginal) contingency tables implied by the data are zero); (2) if $\alpha \to 0$, a modified BIC employing *EDF* (instead of DF) is obtained. To improve statistical inference, various heuristics for adjusting DF in the presence of zero-cell-counts have been devised, see e.g. [2]. The asymptotic expansion in Eq. 9 may motivate the use of EDF as such a heuristics when given *finite* data sets. Note that the EDF, as defined in this paper, are *not* necessarily positive (if EDF are negative, surprising behavior can occur in Eq. 9).

A.2 Proof of Proposition 1 (Limit $\alpha \to 0$)

In the Bayes factor given in Eq. 5, terms cancel out if there are zero-cell-counts:

$$\begin{aligned}
&\log \frac{p(D|m^+)}{p(D|m^-)} \\
= &\sum_{\pi} I(N_\pi) \cdot \log\Gamma(N_\pi + \alpha_\pi) - \sum_{a,\pi} I(N_{a,\pi}) \cdot \log\Gamma(N_{a,\pi} + \alpha_{a,\pi}) \\
- &\sum_{b,\pi} I(N_{b,\pi}) \cdot \log\Gamma(N_{b,\pi} + \alpha_{b,\pi}) + \sum_{a,b,\pi} I(N_{a,b,\pi}) \cdot \log\Gamma(N_{a,b,\pi} + \alpha_{a,b,\pi}) \\
- &\sum_{\pi} I(N_\pi) \cdot \log\Gamma(\alpha_\pi) + \sum_{a,\pi} I(N_{a,\pi}) \cdot \log\Gamma(\alpha_{a,\pi}) \\
+ &\sum_{b,\pi} I(N_{b,\pi}) \cdot \log\Gamma(\alpha_{b,\pi}) - \sum_{a,b,\pi} I(N_{a,b,\pi}) \cdot \log\Gamma(\alpha_{a,b,\pi}) \qquad (10)
\end{aligned}$$

where $I(\cdot)$ is an indicator function (cf. Def. 1). The first four terms approach a constant as $\alpha \to 0$, and we are left with the last four terms. Assuming $\alpha_{a,b,\pi} = \alpha \cdot p(a,b,\pi)$ (cf. Eq. 3), and approximating $1/\Gamma(x) = x + \mathcal{O}(x^2)$ for small x [1], the last four terms in Eq. 10 equal:

$$\begin{aligned}
&\log \frac{\prod_{a,\pi} \Gamma(\alpha_{a,\pi})^{I(N_{a,\pi})} \prod_{b,\pi} \Gamma(\alpha_{b,\pi})^{I(N_{b,\pi})}}{\prod_{a,b,\pi} \Gamma(\alpha_{a,b,\pi})^{I(N_{a,b,\pi})} \prod_{\pi} \Gamma(\alpha_\pi)^{I(N_\pi)}} \\
&= \log \frac{\prod_{a,b,\pi} \alpha_{a,b,\pi}^{I(N_{a,b,\pi})} \prod_{\pi} \alpha_\pi^{I(N_\pi)}}{\prod_{a,\pi} \alpha_{a,\pi}^{I(N_{a,\pi})} \prod_{b,\pi} \alpha_{b,\pi}^{I(N_{b,\pi})}} + \mathcal{O}(\alpha) \\
&= \log[\alpha^{d_{\text{EDF}}} \cdot \frac{\prod_{a,b,\pi} p(a,b,\pi)^{I(N_{a,b,\pi})} \prod_{\pi} p(\pi)^{I(N_\pi)}}{\prod_{a,\pi} p(a,\pi)^{I(N_{a,\pi})} \prod_{b,\pi} p(b,\pi)^{I(N_{b,\pi})}}] + \mathcal{O}(\alpha) \qquad (11)
\end{aligned}$$

Independently of a particular prior choice – as long as $p(\cdot)$ is strictly positive – the value of the ratio is positive and finite. Now, as $\alpha \to 0$, we have $\log \frac{p(D|m^+)}{p(D|m^-)} \to -\infty$ if $d_{\mathrm{EDF}} > 0$, and $\frac{p(D|m^+)}{p(D|m^-)} \to +\infty$ if $d_{\mathrm{EDF}} < 0$. Moreover, a strictly positive prior over the network structures entails finite values of the prior ratios and hence the limiting behavior applies also to the posterior ratio. *qed.*

When $d_{\mathrm{EDF}} = 0$, it is indeed true that the value of the log Bayes factor can converge to any (possibly finite) value as $\alpha \to 0$. It is given by: (1) the first four terms in Eq. 10 when $\alpha \to 0$, which approximate the BIC with $d_{\mathrm{EDF}} = 0$, and hence the maximum likelihood ratio, for large N (cf. also Appendix A.1); plus (2) the last line in Eq. 11, which depends only on the prior over the pseudo-counts, p, since $\alpha^{d_{\mathrm{EDF}}} = 1$. The value of the Bayes factor can be therefore easily set by adjusting the prior weights $p(a, b, \pi)$.